从概念识别到自动化测量：基于大语言模型的国家刻板印象评估

王艺霖[1,2]　赵楠[1,2]　朱廷劭[1,2*]
1 中国科学院心理研究所行为科学重点实验室　北京　100101
2 中国科学院大学心理学系　北京　100049

摘要：本研究以国家刻板印象为例，探索了一种基于大语言模型的心理指标评估方法，实现从概念识别直接到自动化测量的完整流程。研究一基于大语言模型提取自由描述文本中的国家刻板印象内容，并结合文本挖掘方法，再次通过大语言模型归纳出国家刻板印象的跨文化核心维度；研究二进一步基于大语言模型，构建了国家刻板印象自动化维度测量模型，并检验了模型的性能。结果显示：（1）大语言模型揭示了国家刻板印象的五个跨文化核心维度：文化丰富性、发达进步、强权威胁、社会平等、专制独裁；（2）基于大语言模型的自动化维度具有良好的效标效度和重测信度。本研究提出了大语言模型评估（LLM Rating)方法，为心理学评估提供了一种新的跨学科角度，展示了大语言模型在社会认知研究中的潜在应用价值。

关键词：大语言模型，国家刻板印象，社会认知，心理测量

From Concept Identification to Automated Measurement: Assessing National Stereotypes with Large Language Models

Abstract: This study explores a large language model-based psychological assessment method using national stereotypes as a case study, achieving a complete process from concept identification to automated measurement. Study 1 employed large language models to extract national stereotype content from free-description texts and, in combination with text mining methods, further utilized large language models to identify the cross-cultural core dimensions of national stereotypes. Study 2 developed an automated dimension measurement model based on large language models and evaluated its performance. The results indicate that: (1) large language models revealed five cross-cultural core dimensions of national stereotypes: cultural richness, development and progress, dominance and threat, social equality, authoritarianism and dictatorship; (2) the automated dimension measurement based on large language models demonstrated high criterion validity and test-retest reliability. This study proposes LLM rating, which provides a new interdisciplinary perspective for psychological assessment and highlights the potential applications of large language models in social cognition research.

Key Words: Large Language Model, National Stereotype, Social Cognition,

* 通讯作者：朱廷劭; tszhu@psych.ac.cn

Psychological Assessment

* 通讯作者：朱廷劭; tszhu@psych.ac.cn

1 引言

国家刻板印象深刻影响着人们对与其他国家的态度、行为以及国际间的交流与合作 (Esses et al., 2012; Herrmann et al., 1997; Mercer, 2018)。准确评估国家刻板印象不仅有助于理解群体认知模式，还对政策制定、国际关系及全球市场战略具有重要意义 (Herrmann et al., 1997; Mercer, 2018)。然而，如何高效、客观地评估国家刻板印象，对于促进跨文化理解与合作、针对性地调整战略方针具有重要意义。

在国家刻板印象的维度定义上，早期的刻板印象内容模型提出了“温暖—能力”双维度框架 (Fiske et al., 2002)，但该模型主要针对社会群体，而非国家层面的刻板印象。后续研究在国家形象领域扩展了这一框架，例如 Martin 和 Eroğlu (1993)提出政治、经济和技术三维度，Linssen 和 Hagendoorn (1994)以及 Poppe 和 Linssen (1999)在此基础上增加了文化与地理特征（如国土面积）。基于这些维度的相关概念，研究者可以开展后续一系列的测量，例如构建量表量化被试对于某一国家的印象 (Buhmann & Ingenhoff, 2015)。这种从概念定义到量表测量的评估过程，往往涉及多个步骤，包括维度的主观定义、测量工具的信效度验证等，使得操作过程较为复杂且耗时耗力。因此，如何建立更高效、稳健的国家刻板印象评估方式，成为亟待解决的问题。

随着自然语言处理技术的兴起，研究者们开始尝试使用大语言模型来捕捉文本内容中有关刻板印象的丰富内容。研究表明，大语言模型通过基于大规模语料库的预训练，能够捕捉语言中的隐含偏见 (X. Bai et al., 2024; Wang & Lin, 2024)。大语言模型不仅可以识别这些心理现象，还能实现对自动化测量。在对自由描述等开放式文本进行评估时，大语言模型可以作为测量工具，模拟人类专家对内容的评判行为 (Chen et al., 2024; Xiao & Yang, 2024; Zhu et al., 2023)。例如 Huang 等人 (2024)通过调用大语言模型对被试的自由描述文本进行分析，得到了关于其生活满意度的维度评分，这些评分与人类专家较为一致。然而，这些使用大语言模型进行自动化测量的方法仍然沿用了传统心理学的研究框架，即依赖于已有量表构建的维度及其定义，使得评分结果可能包含量表构建时存在的测量误差。此外，这种方法没有利用好大语言模型在识别丰富语义概念和客观量化评分上的不同优势，因而没有充分发挥大语言模型在心理指标评估中的潜力。

因此，本研究利用大语言模型提出了一种新的研究范式，不再依赖传统心理学量表的构建过程，而是充分发挥大语言模型在概念识别和量化评分方面的优势，通过两步实现国家刻板印象的自动化测量：（1）结合其他文本挖掘技术，利用大语言模型从自由描述文本中提取国家刻板印象的维度；（2）基于这些维度，利用

多种大语言模型构建自动化评分模型，对自由描述文本实现批量评估，同时验证这种自动化评估方式的性能。据此，本研究提出如下假设：（H1）国家刻板印象的维度不局限于经典的刻板印象内容模型的两个维度，也会在前人关于国家印象维度的基础上有重叠与延伸；（H2）通过结合大语言模型，可以构建模拟人类专家判断的国家刻板印象维度评分模型，且模型信效度表现优异。

2 研究一 国家刻板印象的跨文化核心维度

2.1 被试

为了对比不同文化背景下被试对他国持有的刻板印象，我们考虑招募中国和美国的被试。尽管两国在国际地位上实力比较一致，在全球性议题中均承担着相似的责任与影响力；但两国在社会、政治等多方面截然不同 (Pu, 1989)。其次，英文和中文均为世界上最广泛使用的语言之一，具有全球性的传播力与影响力。两种语言在全球文化交流中扮演着重要角色，是文化传播和全球沟通的桥梁 (“List of Languages by Number of Native Speakers,” 2024)。因此，收集来自这两个国家的被试数据，不仅具有跨文化研究的价值，还能够反映出全球范围内可能存在的刻板印象模式。根据预实验的功效分析，我们最终将每个国家的目标样本量初步定为 197 名被试（预实验详见补充材料）。

2024 年 9 月，我们通过问卷星平台基于其在线程序招募了 197 名中国被试：（1）年龄为 18 岁或以上；（2）母语为中文；（3）居住在中国；（4）学历为高中或以上。针对招募的数据我们设定了两个排除标准：（1）若内容的汉字字数少于 50 字，则该内容将被排除；（2）被试排除标准：若被试未通过态度认真考量或不符合招募标准，则其所有数据将被排除。根据上述标准，共有 6 名被试和 19 条内容被排除。最终，样本中保留了 191 名被试（83 名男性，107 名女性，1 名其他性别；年龄：$M = 31.28$ 岁，$SD = 16.10$），共计 375 条有效内容。

2024 年 10 月，我们通过 CloudResearch Connect 平台基于 JavaScript 编写的在线程序招募了 197 名美国被试，招募标准为：（1）年龄为 18 岁或以上；（2）母语为英文；（3）居住在美国；（4）学历为高中或以上。针对招募的数据我们设定了两个排除标准：（1）若内容的单词数少于 50 词，则该内容将被排除；（2）被试排除标准：若被试未通过态度认真考量或不符合招募标准，则其所有数据将被排除。根据上述标准，共有 21 名被试和 55 条内容被排除。最终，样本中保留了 176 名被试（91 名男性，85 名女性；年龄：$M = 47.08$ 岁，$SD = 16.09$），共计 339 条有效内容。

2.2 刺激材料

考虑到中美关系在国际局势中的重要性 (Kissinger, 2012)，我们在本研究中着重探讨这两个国家公民对彼此国家持有的刻板印象。为了拓展跨文化比较的深度和广度，我们还选择了印度作为对比国家，印度在一些方面与中美的相似性为国家刻板印象研究提供了一个有价值的参考。

对于呈现给中国被试的国家，我们选择了美国和印度这两个国家作为刺激材料。就美国而言，我们考虑到的是中美在世界舞台上匹敌的国际地位，以复杂动态的双边关系 (Yan, 2010)。对于印度，则考虑到的是其与中国在某些关键方面的相似性，以及同样复杂动态的双边关系 (Yu et al., 2022)。

对于呈现给美国被试的国家，我们按照第 3 章"3.3.1　刺激材料"的理论依据选择了中国和印度这两个国家作为刺激材料。这一选择不仅基于中美和美印之间复杂的国际关系，还考虑到了这三个国家在全球政治、经济及文化等多个领域的显著地位 (Curtis, 2008)。

2.3 实验流程

中国被试的实验通过问卷星平台的在线程序开展，美国被试的实验通过基于 JavaScript 编写的在线程序开展。每名被试需要完成两个试次。在每个试次中，被试需对呈现的国家进行自由描述任务。相同国家只呈现一次，不同国家的呈现顺序以随机化方式确定。在每个试次中，被试被明确告知需站在其所属国家公民的角度，对给定国家形成心理表征，并撰写一篇类似短文的自由描述。撰写的文字内容应该基于当下的现实生活，不应包含任何童话、科幻、过去的印象、尚未发生的事件等内容。

为最大程度地激发被试的即时反应并促进自由描述，每个试次的答题时间被限制为 5 分钟，要求被试在规定时间内尽可能多地撰写，同时尽量少地修改内容。

为避免被试因连续书写产生疲劳，在两个试次之间设置了没有时间限制的休息界面。为了探究受众特征对其所持有的国家刻板印象的影响，在完成所有试次后，被试需要填写一份简短的人口学信息问卷。

2.4 分析方法

本研究首先使用大语言模型对中美被试群体的自由描述进行国家刻板印象内容的提取，然后运用网络分析和主题建模等多种文本挖掘方式解释其中蕴藏的潜在结构，最后再次运用大语言模型对这些潜在结构进行命名和整合，从而得到国家刻板印象的通用维度模型。

参考 Wang 和 Lin (2024)的分析方法，我们首先使用大语言模型中的 GPT-4o

(OpenAI, Hurst, et al., 2024)对每个自由描述进行文本预处理，以纠正其中的语法错误。随后，我们设计了精确的提示词，要求在伦理准则方面较为不受限制的 GPT-3.5-turbo 从每个自由描述中识别与提示词中指定国家相关的刻板印象词语(OpenAI, Achiam, et al., 2024)。针对每个自由描述，我们最终获得了一组能够反映国家刻板印象特征的词语。最终，中国被试对应 2488 个词语，其中无重复词语为 1542 个；美国被试对应 2471 个词语，其中无重复词语为 1331 个。为了去掉噪音，只纳入了上述至少在 2%的被试自由描述中出现过的国家刻板印象词语（中国被试共计 375 个自由描述，美国被试共计 339 个自由描述），这样在中国被试中就剩下了 36 个无重复的国家刻板印象词语；在美国被试中就剩下了 62 个无重复的国家刻板印象词语。针对中美被试，对于每个保留的国家刻板印象词语，我们用该词在自由描述中的出现次数除以该自由描述中提取到的国家刻板印象总词数，从而计算出其在每个自由描述中的出现比率。这样，我们就能构建国家刻板印象词频矩阵，其中每行表示每个自由描述，每列表示保留得到的国家刻板印象词语

针对上述得到的国家刻板印象词频矩阵，在网络分析中，根据每对国家刻板印象词语在自由描述中的出现比率，我们计算出词语之间的 Spearman 相关系数矩阵。基于相关系数矩阵，我们将每个刻板印象词语表示为一个节点，这些词语之间的 Spearman 相关系数的绝对值则表示为连接这些节点的边的权重。然后，我们进行了网络分析（network analysis）来推导国家刻板印象存在的维度表征。我们使用 Louvain 算法 (Blondel et al., 2008)对网络进行社区检测，识别具有更强内部关联的刻板印象词语群体（即潜在的聚类）。在主题建模（topic modeling）分析中，为了从国家刻板印象词语中提取潜在主题，我们采用了三种常见的主题建模方法：潜在狄利克雷分配（Latent Dirichlet Allocation, LDA）、非负矩阵分解（Non-negative Matrix Factorization, NMF）和潜在语义分析（Latent Semantic Analysis, LSA）。

对于上述两种分析方法（网络分析、主题建模）得到的潜在结构，我们使用 GPT-4o 对每种方法得到的国家刻板印象结构进行命名，命名依据为代表性词语及其重要性程度，并为每个名称提供解释。

针对上述不同分析方法得到的国家刻板印象潜在结构（包含维度名称及其定义），我们使用 GPT-4o 依据文本嵌入的量化分析，对中美被试数据基于所有分析方法得到的潜在结构进行整合。

2.5 结果

在网络分析中，对于中国被试，提取了六个社区，其中，国家刻板印象词语的节点强度中心性平均值为 1.46，最小值为 0.30，最大值为 2.59。对于美国被试，

网络分析提取了五个社区，其中，国家刻板印象词语的节点强度中心性平均值为 2.28，最小值为 0.94，最大值为 4.04。表 1 列出了网络分析得到的中美被试所持有的他国刻板印象结构，并展示了通过 GPT-4o 命名的各个社区及其含义，以及每个社区的代表性国家刻板印象词语。

表 1 基于网络分析的中美被试持有的他国刻板印象结构

被试	社区	含义	代表性词语
中国	经济差距	该国家在经济发展和资源分配上的不平衡印象	贫富、落后、丰富
	强国形象	经济发达、政治民主且国力强大的国家形象	发达、民主、强大
	社会不平	社会中存在的不平等现象	种族歧视、种姓制度、贫富差距
	不安因素	对不确定性和潜在威胁的感知	危险、发展、歧视
	科技强国	该国家在科技领域的领先地位以及由此可能带来的社会不平等问题	科技、领先、不平等
	多元发展	文化、开放和经济三个方面的综合影响，体现了一个国家在多方面的进步与包容性	文化、开放、经济
美国	负面认知	将一个国家视为拥有强大军事力量的共产主义和威权主义威胁，且评价负面	communist, authoritarian, threat, negative, military
	文化丰富性	认为一个国家充满活力、多样化，并且强调传统和丰富多彩的元素	exotic, crowd, diverse, traditional, colorful
	负面刻板印象	认为某个国家面临贫穷和人口过剩问题，但也承认其勤奋和智慧等积极特质	dirty, poor, overpopulate, hardworking, intelligent
	印度	反映了关于印度的常见刻板印象，突出其文化丰富性、宝莱坞影响力、经济挑战、技术进步以及尊重社会价值	culture, bollywood, poverty, technology, respect
	进步	认为一个国家在经济上先进、技术上创新，并在人权方面提供积极的机会	economic, tech, positive, opportunity, human-rights

在主题建模中，依据困惑度（perplexity）、重构误差（reconstruction error）、解释方差（explained variance）等指标与主题数量的关系，我们针对每种主题建模分析方法，不论是中国被试还是美国被试，最终确定的最佳数量都为四个。

在 LDA 分析方法下，对于中国被试，得到的国家刻板印象词语的权重平均值为 0.82，最小值为 0.25，最大值为 5.79。对于美国被试，得到的国家刻板印象词语的权重平均值为 0.71，最小值为 0.25，最大值为 5.31。表 2 列出了 LDA 得到的中美被试所持有的他国刻板印象结构，并展示了通过 GPT-4o 命名的各个主题及其含义，以及每个主题的代表性国家刻板印象词语。

表 2 基于 LDA 的中美被试持有的他国刻板印象结构

被试	主题	含义	代表性词语
中国	落后国家	对某些国家经济和环境状况的负面刻板印象	落后、脏乱、贫穷
	发展霸权	对国家在追求发展的同时可能涉及霸权主义行为的复杂看法	发展、霸权主义、复杂
	经济差距	对国家经济实力和发展程度的刻板印象	贫富、强大、发达

	自由民主	对某些国家在政治和社会制度上追求自由和民主的刻板印象，同时也可能涉及其在国际事务中的霸权地位	自由、霸权、民主
美国	人口过多的担忧	认为某些国家人口过多，构成威胁，同时具有智慧、宗教信仰和贫困的特征	overpopulate, threat, intelligent, religious, poor
	文化刻板印象	认为某些国家是共产主义的、异域风情的，并以廉价制造业闻名，常被视为既美丽又经济实惠	communist, exotic, cheap, beautiful, manufacturing
	发展中国家的刻板印象	对贫穷、文化丰富性以及拥挤和清洁等挑战的看法	poor, culture, dirty, overcrowd, diverse
	社会经济认知	与经济状况、人口密度和文化特征相关的刻板印象，影响对智慧和生活方式的看法	poverty, crowd, control, smart, spicy

在 NMF 分析方法下，对于中国被试，得到的国家刻板印象词语的权重平均值为 0.06，最小值为 0，最大值为 1.04。对于美国被试，得到的国家刻板印象词语的权重平均值为 0.05，最小值为 0，最大值为 1.10。表 3 列出了 NMF 得到的中美被试所持有的他国刻板印象结构，并展示了通过 GPT-4o 命名的各个主题及其含义，以及每个主题的代表性国家刻板印象词语。

表 3 基于 NMF 的中美被试持有的他国刻板印象结构

被试	主题	含义	代表性词语
中国	自由发展	对一个国家在自由度和经济发展方面的刻板印象	自由、发达、发展
	脏乱贫穷	对某些国家环境和经济状况的负面刻板印象	脏乱、落后、贫穷
	贫富差距	主要关注国家间经济不平等的问题，尽管环境和科技也有一定影响	贫富、环境、科技
	强权	对国家力量和影响力的刻板印象，强调其强大和霸权的特征	强大、霸权、合作
美国	贫穷	认为某些国家在经济上处于劣势，与恶劣的生活条件和高人口密度相关联	poor, dirty, overpopulate, religious, overcrowd
	社会经济挑战	与某些国家经济困难、文化方面和环境问题相关的刻板印象	poverty, culture, overcrowd, bollywood, pollution
	共产主义威胁	认为某些国家是共产主义国家，重点是它们被视为构成威胁、控制资源和军事存在	communist, threat, cheap, control, military
	文化认知	对国家作为异域且多样的认知，融合了传统元素与权力感，通常与拥挤的环境联系在一起	exotic, diverse, crowd, traditional, powerful

在 LSA 分析方法下，对于中国被试，得到的国家刻板印象词语的权重平均值为 0.03，最小值为–0.40，最大值为 0.81。对于美国被试，得到的国家刻板印象词语的权重平均值为 0.01，最小值为–0.56，最大值为 0.60。表 4 列出了 LSA 得到的中美被试所持有的他国刻板印象结构，并展示了通过 GPT-4o 命名的各个主题及其含义，以及每个主题的代表性国家刻板印象词语。

表 4 基于 LSA 的中国被试持有的他国刻板印象结构

被试	主题	含义	代表性词语
中国	社会差距	对国家经济不平等、自由度和发展水平的刻板印象	贫富、自由、发达
	脏乱贫穷	一个国家在环境卫生、经济发展和生活条件上的负面刻板印象	脏乱、落后、贫穷
	贫富差距	对国家经济不平等和社会分层的刻板印象	贫富、强大、环境
	强国形象	对一个国家的强大和发达的印象，并且强调了合作的重要性	强大、发达、合作
美国	社会经济挑战	与经济困难、生活条件和与贫困及人口过剩相关的文化方面刻板印象	poor, poverty, dirty, overpopulate, culture
	贫困刻板印象	关于被认为贫穷国家的普遍刻板印象，包括与缺乏清洁、高宗教信仰、人口过剩以及独特饮食特征相关联的看法	poor, dirty, religious, overpopulate, spicy
	社会经济状况	反映了与贫困和文化认知相关的刻板印象，包括对某些国家过度拥挤和异域风情的看法	poverty, culture, overcrowd, exotic, crowd
	文化认知	对一个国家作为异域且多样化的认知，融合了传统元素与权力感，通常与人们如何看待文化丰富性和影响力相关联	exotic, crowd, diverse, traditional, powerful

基于中美被试在所有分析方法中得到的潜在结构，我们使用 GPT-4o 依据文本嵌入的量化分析，对这些潜在结构进行维度总结。得到的跨文化维度模型见表 5。

表 5 GPT-4o 整合的国家刻板印象通用维度模型

维度	含义
文化丰富性	指对一个国家的文化多样性、全球影响力、精神意义以及对文化交流的开放程度的认知，体现该国文化的吸引力与包容性
发达进步	指对一个国家的经济发展、科技进步、资源丰富及现代化的认知，强调其发展的多样性与能力
强权威胁	指对一个国家作为全球强权的认知，体现为其军事力量、资源垄断及对其他国家的控制，并伴随对全球稳定的威胁或国际关系中的竞争
社会平等	指对一个国家在资源分配公平性、减少社会阶层差异方面的认知
专制独裁	指对一个国家受严格集权控制，限制自由，并维护传统意识形态的认知

3 研究二 国家刻板印象的自动化测量

3.1 被试、刺激材料与实验流程

针对研究一保留的所有有效美国被试，我们于 2024 年 11 月再次通过 CloudResearch Connect 平台向他们发出邀请，最终招募了 59 名美国被试。招募数据的排除标准同研究一中的美国被试。根据上述标准，共有 4 条内容被排除。最终，样本中保留了 59 名被试（30 名男性，29 名女性；年龄：M = 47.29 岁，SD = 17.34），共计 114 条有效内容。

刺激材料同研究一中呈现给美国被试的实验材料，即中国和印度；实验流程也与研究一的美国被试实验流程相同。

3.2 分析方法

本研究采用了 GPT-4o (OpenAI, Hurst, et al., 2024)、DeepSeek-R1 (DeepSeek-AI et al., 2025)、Llama 3.3 (Grattafiori et al., 2024)和 Qwen-max (J. Bai et al., 2023)这四种广泛应用且性能领先的大语言模型分别构建国家刻板印象维度评分模型。

在本研究的模型调用方式上，GPT-4o 直接通过 OpenAI 平台的 API key 进行调用，而 DeepSeek-R1、Llama 3.3 和 QWen-max 则是通过阿里云百炼平台的 API key 进行调用。由于阿里云百炼在内容审核上实施严格的敏感词检测机制，因此在通过该平台调用模型前，我们对文本内容进行了预处理，以确保文本能够顺利通过平台的审核。我们对文本内容的预处理操作作为：（1）将“China”（不区分大小写）统一替换为“country”；（2）对具体政治人物的姓名进行了删除。

模型的构建与运行主要包括以下三个步骤：首先，通过人类专家评分建立基准数据，为后续的自动化评分模型提供参考标准。其次，设定大语言模型的提示词，以确保模型能够准确理解国家刻板印象的五个维度，并在评分过程中尽可能贴近专家的判断。最后，运行基于大语言模型的国家刻板印象评分模型，对不同自由描述文本内容进行评分。以下将详细介绍各个步骤的具体实施过程。

我们从 59 名有效被试中随机选择 50 名被试，整合这 50 名被试在 2024 年 10 月收集的数据（即研究一的数据）和 2024 年 11 月的数据（即研究二的数据），共计 200 条针对中国和印度的内容，进行后续的重测信度分析。

对于本研究招募得到的 114 条有效内容，排除用作重测信度分析的 50 名被试对应的 100 条有效内容，最终剩余 14 条有效内容。同时，我们从美国网站收集了 20 条与国家印象相关的文本内容。基于研究一美国被试自由描述平均词数（100 词），我们设定网络信息收集标准：每条内容需介于 90-100 词（包括 909 和 100）。最终，将有效内容和网络信息合并，得到 34 条内容用于后续分析。

随后，三名英文流利的专家依据研究一的国家刻板印象通用模型，对这 34 条内容在五个维度上按照–2（完全相反）到 2（完全相同）进行五点评分。评分前，专家接受统一培训，详细了解每个维度及其定义。评分时，专家需讨论并达成一致后给出最终分数。

我们从这 34 条内容中随机选取了在每个国家刻板印象维度上都有体现的 4 条内容，作为大语言模型提示词中的示例，以展示人类专家的评分规律和准则。其余 30 条用于后续效标效度分析。为确保大语言模型评分的拟人化，提示词包含与专家评分英文指导手册一致的信息。最终，我们基于专家评分示例和指导手册内容设定提示词。

考虑到再次参与实验的被试可能针对同一国家撰写不同主题的描述（如先描述中国经济，后描述中国文化），而重测信度评估需确保测量内容一致性，我们对用于重测信度分析的 200 条针对中国和印度的内容进行了清洗，剔除主题明显

不一致的自由描述。最终，保留 31 名被试提供的 94 条有效内容。

在运行基于大语言模型的国家刻板印象维度评分模型时，我们为每个模型（GPT-4o、DeepSeek-R1、Llama 3.3、QWen-max）单独部署评分模型，确保在相同提示词和评分准则下独立评分。由于不同平台对文本处理要求不同，GPT-4o 直接评分，而 DeepSeek-R1、Llama 3.3 和 QWen-max 基于预处理文本评分。最终，每个模型分别对 30 条效标效度分析内容和 94 条重测信度分析内容，按国家刻板印象的五个维度独立评分。

在效标效度检验上，为了独立评估每个基于大语言模型的评分模型在国家刻板印象维度评分任务中的有效性，我们对其均进行了 Spearman 相关性分析，从中比较 30 条内容上大语言模型评分与人类专家评分之间的相关程度。具体而言，在每个基于大语言模型的评分模型中，我们对所有内容在每个维度上的得分进行了相关性分析。由于我们计算效标效度时，分别对四个大语言模型中的每个维度都进行了相关性分析，因此我们采用 Holm-Bonferroni 方法对 20 次相关性分析带来的多重比较结果进行了校正。

在重测信度检验上，为了独立评估每个基于大语言模型的评分模型生成的评分是否在时间上具有一定的稳定性，我们对其均进行了 Spearman 相关性分析，从中比较 94 条内容上构建的国家刻板印象维度评分模型在跨时间上的一致性。具体而言，在每个基于大语言模型的评分模型中，针对每名被试就同一国家的前后自由描述在同一维度上的得分，我们进行了前后得分的相关性分析。由于我们计算重测信度时，分别对四个大语言模型中的每个维度都进行了相关性分析，因此我们采用 Holm-Bonferroni 方法对这 20 次相关性分析带来的多重比较结果进行校正。

3.3 结果

在国家刻板印象的五个维度上，基于不同大语言模型的评分模型的效标效度均表现良好（Spearman 相关系数范围为 .47 到 .86，测量均值与效标均值之差的绝对值范围为 0.00 到 0.21，校正后的 p 值范围为 3.23×10^{-8} 到 .01）。不同评分模型在国家刻板印象维度上的效标效度见表 6。

表 6 国家刻板印象维度评分模型的效标效度（r）

维度	GPT-4o	DeepSeek-R1	Llama 3.3	QWen-max
文化丰富性	$.86^{***}$	$.86^{***}$	$.84^{***}$	$.77^{***}$
发达进步	$.76^{***}$	$.75^{***}$	$.73^{***}$	$.70^{***}$
强权威胁	$.74^{***}$	$.67^{***}$	$.51^{*}$	$.63^{**}$
社会平等	$.81^{***}$	$.84^{***}$	$.76^{***}$	$.80^{***}$
专制独裁	$.60^{**}$	$.60^{**}$	$.47^{*}$	$.50^{*}$

注： $*p < .05$，$**p < .01$，$***p < .001$，$^{a}p < .1$。

在国家刻板印象的五个维度上，基于不同大语言模型的评分模型的重测信度均表现良好（Spearman 相关系数范围为 .52 到 .84，前后两次测量均值之差的绝对值范围为 0.00 到 0.19，校正后的 p 值范围为 2.56×10^{-12} 到 1.69×10^{-4}）。不同评分模型在国家刻板印象维度上的重测信度见表 7。

表 7 国家刻板印象维度评分模型的重测信度（r）

维度	GPT-4o	DeepSeek-R1	Llama 3.3	QWen-max
文化丰富性	$.72^{***}$	$.63^{***}$	$.78^{***}$	$.62^{***}$
发达进步	$.58^{***}$	$.68^{***}$	$.78^{***}$	$.72^{***}$
强权威胁	$.83^{***}$	$.82^{***}$	$.83^{***}$	$.52^{***}$
社会平等	$.67^{***}$	$.59^{***}$	$.74^{***}$	$.68^{***}$
专制独裁	$.84^{***}$	$.69^{***}$	$.61^{***}$	$.54^{***}$

4 讨论

本研究融合计算机科学与心理学，探索基于大语言模型的国家刻板印象评估方法，构建了从概念识别到自动化测量的社会认知计算框架。研究一首先利用大语言模型提取自由描述中的国家刻板印象，并结合多种文本挖掘技术揭示其潜在结构，进而再次利用大语言模型量化整合出跨文化的国家刻板印象维度模型。结果展示出五个普适性维度：文化丰富性、发达进步、强权威胁、社会平等和专制独裁。研究二进一步验证了大语言模型在心理指标自动化测量中的效度与信度，结果显示其评分与人类专家高度一致，并在一定程度上保持跨时间稳定性。这表明大语言模型能够完成从概念识别到测量的自动化评估过程，提高分析效率和客观性。

以往跨学科的心理测量方式通常依赖于“概念定义—量表构建—评估模型构建—自动化测量”路径。即研究者需先基于理论定义概念及维度，再开发相应的心理量表，并通过大规模收集被试数据，进而训练模型实现从行为数据到心理指标的映射，最后得以开展自动化测量 (Li et al., 2014; Wang et al., 2020)。然而，这一过程费时费力且易累积测量误差，最后实现的自动化测量的准确性在很大程度上取决于量表开发的质量。本研究以国家刻板印象为例，提出了一种新的心理指标评估方式，即“大语言模型一体评估路径”，直接从识别概念到进行自动化测量，避免因量表构建误差导致的偏差，以及时间和金钱成本。

与假设 1 一致，研究一发现，国家刻板印象具有多维结构，这也在以往的研究中得到验证。例如文化丰富性维度契合国家形象四维模型中的审美维度 (Buhmann & Ingenhoff, 2015)，发达进步维度反映 Martin 和 Eroğlu (1993)提出的经济技术因素，强权威胁维度与刻板印象内容模型中的高能力-威胁认知一致

(Fiske et al., 2007)，社会平等维度被证实关乎国家的国际声誉 (Wilkinson & Pickett, 2011)，而专制独裁维度则往往与负面国家评价相关 (Inglehart & Welzel, 2005; Pratto et al., 2006)。

基于研究一的发现，研究二构建了国家刻板印象维度自动化评分模型，并验证其效标效度与重测信度。与假设 2 一致，我们发现所有大语言模型在国家刻板印象维度自动化评分的能力上表现出显著的可靠性和稳定性。在效标效度上，尤其是在较直观的维度（如文化丰富性）上，其表现高度一致；但在较抽象的维度（如专制独裁）上，与人类评分的一致性则较低 (Dancey & Reidy, 2007)。这种差异可能源于专制独裁等维度通常包含隐喻（例如“铁腕”）、对情境的高度依赖（例如“地方权力自主”可能与“自由民主的政治”含义相似）、褒贬交织（例如“社会井然有序”可能与“专制”同时出现）等复杂特性。这表明，大语言模型在需要更复杂推理的维度上，其评分表现可能受隐喻、上下文依赖等因素影响 (Anagnostidis & Bulian, 2024)。在重测信度上，较为稳定、客观的维度（如强权威性）的跨时间稳定性较高，而感知较为主观的维度（如专制独裁）的跨时间稳定性则较低 (Dancey & Reidy, 2007)。具体来说，强权威胁维度通常涉及国家的军事力量、全球政治影响力等较为客观、固定的属性，不太容易改变，因此大语言模型在不同时间点上对同一国家的描述能保持较高的一致性 (Waltz, 2010)。而专制独裁维度往往涉及政策方针的调整，评价更为主观，并可能受到文本中细微语境变化的影响，导致大语言模型在不同时间点上的评分略有波动 (Dahl, 2008)。

尽管本研究证明了大语言模型在心理指标测量，尤其是国家刻板印象评估中的可行性，但仍然存在一定局限性。首先，尽管样本规模较大且具有代表性，但在跨文化背景下，我们的研究仍有进一步扩展的空间。未来研究可以涵盖更多国家，以验证本研究归纳出的国家刻板印象核心维度的跨文化普适性 (D. Martin et al., 2014)。其次，我们使用了网络分析和多种主题建模方法提取国家刻板印象的潜在结构，并最终通过大语言模型基于文本嵌入的量化方式进行整合。然而，不同方法的结果可能受到各自算法假设的影响，这种多方法整合虽然增强了研究结果的可靠性 (Gerlach et al., 2018)，但可能掩盖了某些方法独特的贡献。未来研究可以采用其他新兴的文本分析方法，例如深度学习模型，以验证这些跨文化核心维度的鲁棒性。此外，尽管构建的模型在效标效度和重测信度上各个维度都达到显著相关，但有的相关系数并没有达到强相关水平，尤其是在某些依赖于微妙语义或特定语境的维度中，这可能与大语言模型对语义细节的解析能力，以及维度本身定义的稳定性有关。未来研究可以通过引入多模态提示词或上下文信息来增强模型对复杂语境和维度定义随时间变化的理解 (Anagnostidis & Bulian, 2024)。

5 结论

本研究以国家刻板印象为例，提出了一种基于大语言模型的心理指标评估方法，即大语言模型评估（LLM Rating），实现了从概念识别直接到自动化测量的一体化评估。通过大语言模型的语义理解与测量能力，本研究自动提取了国家刻板印象的核心维度，并在此基础上构建了性能良好的自动化维度测量模型。这种方式突破了传统心理学研究中需要构建心理量表再评分的局限性、减少了累积误差，为社会认知研究提供了一种更具灵活性和可扩展性的评估角度。

参考文献

Anagnostidis, S., & Bulian, J. (2024). *How susceptible are LLMs to influence in prompts?* (No. arXiv:2408.11865). arXiv. https://doi.org/10.48550/arXiv.2408.11865

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., … Zhu, T. (2023). *Qwen technical report* (No. arXiv:2309.16609). arXiv. https://doi.org/10.48550/arXiv.2309.16609

Bai, X., Wang, A., Sucholutsky, I., & Griffiths, T. L. (2024). *Measuring implicit bias in explicitly unbiased large language models* (No. arXiv:2402.04105). arXiv. https://doi.org/10.48550/arXiv.2402.04105

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). *Fast unfolding of communities in large networks*. https://doi.org/10.1088/1742-5468/2008/10/P10008

Buhmann, A., & Ingenhoff, D. (2015). Advancing the country image construct from a public relations perspective. *Journal of Communication Management*, *19*(1), 62–80. https://doi.org/10.1108/JCOM-11-2013-0083

Chen, D., Huang, Y., Ma, Z., Chen, H., Pan, X., Ge, C., … Zhou, J. (2024). Data-Juicer: A one-stop data processing system for large language models. In *Companion of the 2024 International Conference on Management of Data* (pp. 120–134). ACM. https://doi.org/10.1145/3626246.3653385

Curtis, L. (2008). U.S.–India relations: The China factor. *Backgrounder*, *2209*.

Dahl, R. A. (2008). *Polyarchy: Participation and opposition*. Yale university press. https://books.google.com/books?hl=zh-CN&lr=&id=JcKz2249PQcC&oi=fnd&pg=PP9&dq=Polyarchy:+Participation+

and+Opposition&ots=PF4DXW_-N2&sig=ZO-KkPMzOckDCPh7RpbH389kS5U

Dancey, C. P., & Reidy, J. (2007). *Statistics without maths for psychology*. Pearson/Prentice Hall.

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., … Zhang, Z. (2025). *DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning* (No. arXiv:2501.12948). arXiv. https://doi.org/10.48550/arXiv.2501.12948

Esses, V. M., Veenvliet, S., & Medianu, S. (2012). The dehumanization of refugees: Determinants and consequences. In S. Wiley, G. Philogène, & T. A. Revenson (Eds.), *Social categories in everyday experience* (pp. 133–150). American Psychological Association. https://doi.org/10.1037/13488-007

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83. https://doi.org/10.1016/j.tics.2006.11.005

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, *82*(6), 878–902. https://doi.org/10.1037/0022-3514.82.6.878

Gerlach, M., Peixoto, T. P., & Altmann, E. G. (2018). A network approach to topic models. *Science Advances*, *4*(7), eaaq1360. https://doi.org/10.1126/sciadv.aaq1360

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., … Ma, Z. (2024). *The Llama 3 herd of models* (No. arXiv:2407.21783). arXiv. https://doi.org/10.48550/arXiv.2407.21783

Herrmann, R. K., Voss, J. F., Schooler, T. Y. E., & Ciarrochi, J. (1997). Images in international relations: An experimental test of cognitive schemata. *International Studies Quarterly*, *41*(3), 403–433. https://doi.org/10.1111/0020-8833.00050

Huang, F., Sun, X., Mei, A., Wang, Y., Ding, H., & Zhu, T. (2024). LLM plus machine learning outperform expert rating to predict life satisfaction from self-statement text. *IEEE Transactions on Computational Social Systems*, 1–8. IEEE Transactions on Computational Social Systems. https://doi.org/10.1109/TCSS.2024.3475413

Inglehart, R., & Welzel, C. (2005). *Modernization, cultural change, and democracy: The human development sequence*. Cambridge University Press.

https://doi.org/10.1017/CBO9780511790881

Kissinger, H. A. (2012). The future of U.S.-Chinese relations: Conflict is a choice, not a necessity essay. *Foreign Affairs*, *91*(2), 44–55.

Li, L., Li, A., Hao, B., Guan, Z., & Zhu, T. (2014). Predicting active users' personality based on micro-blogging behaviors. *PLOS One*, *9*(1), e84997. https://doi.org/10.1371/journal.pone.0084997

Linssen, H., & Hagendoorn, L. (1994). Social and geographical factors in the explanation of the content of European nationality stereotypes. *British Journal of Social Psychology*, *33*(2), 165–182. https://doi.org/10.1111/j.2044-8309.1994.tb01016.x

List of languages by number of native speakers. (2024). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=List_of_languages_by_number_of_native_speakers&oldid=1261007365

Martin, D., Hutchison, J., Slessor, G., Urquhart, J., Cunningham, S. J., & Smith, K. (2014). The spontaneous formation of stereotypes via cumulative cultural evolution. *Psychological Science*, *25*(9), 1777–1786. https://doi.org/10.1177/0956797614541129

Martin, I. M., & Eroglu, S. (1993). Measuring a multi-dimensional construct: Country image. *Journal of Business Research*, *28*(3), 191–210. https://doi.org/10.1016/0148-2963(93)90047-S

Mercer, J. (2018). *Reputation and international politics*. Cornell University Press.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., … Zoph, B. (2024). *GPT-4 technical report* (No. arXiv:2303.08774). arXiv. https://doi.org/10.48550/arXiv.2303.08774

OpenAI, Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., … Malkov, Y. (2024). *GPT-4o system card* (No. arXiv:2410.21276). arXiv. https://doi.org/10.48550/arXiv.2410.21276

Poppe, E., & Linssen, H. (1999). In-group favouritism and the reflection of realistic dimensions of difference between national states in central and eastern European nationality stereotypes. *British Journal of Social Psychology*, *38*(1), 85–102. https://doi.org/10.1348/014466699164059

Pratto, F., Sidanius, J., & Levin, S. (2006). Social dominance theory and the dynamics of intergroup relations: Taking stock and looking forward. *European Review of Social Psychology*, *17*(1), 271–320. https://doi.org/10.1080/10463280601055772

Pu, Z. (1989). A comparative perspective on the United States and Chinese constitutions.

William And Mary Law Review, *30*, 867–880.

Waltz, K. N. (2010). *Theory of international politics*. Waveland Press. https://books.google.com/books?hl=zh-CN&lr=&id=OaMfAAAAQBAJ&oi=fnd&pg=PP2&dq=Theory+of+International+Politics&ots=GN3fPj0DwQ&sig=83F5ajGao8V6NLVK68SY9tRk4AM

Wang, Y., & Lin, C. (2024). *Stereotypes at the intersection of perceivers, situations, and identities: Analyzing stereotypes from storytelling using natural language processing*. OSF. https://doi.org/10.31234/osf.io/dvurm

Wang, Y., Wu, P., Liu, X., Li, S., Zhu, T., & Zhao, N. (2020). Subjective well-being of chinese sina weibo users in residential lockdown during the COVID-19 pandemic: machine learning analysis. *Journal of Medical Internet Research*, *22*(12), e24775. https://doi.org/10.2196/24775

Wilkinson, R., & Pickett, K. (2011). *The spirit level: Why greater equality makes societies stronger*.

Xiao, C., & Yang, B. Z. (2024). *LLMs may not be human-level players, but they can be testers: Measuring game difficulty with LLM agents* (No. arXiv:2410.02829). arXiv. https://doi.org/10.48550/arXiv.2410.02829

Yan, X. (2010). The instability of China–US relations. *The Chinese Journal of International Politics*, *3*(3), 263–292. https://doi.org/10.1093/cjip/poq009

Yu, X., Chang, P. K., & Kho, S. N. (2022). Classifying China-India relationships: Cooperation, competition and conflict. *Higher Education and Oriental Studies*, *2*(6), 16–23. https://doi.org/10.54435/heos.v2i6.82

Zhu, L., Wang, X., & Wang, X. (2023). *JudgeLM: Fine-Tuned large language models are scalable judges* (No. arXiv:2310.17631). arXiv. https://doi.org/10.48550/arXiv.2310.17631